Enfoque de estimación puntual en Análisis de Riesgo Ambiental

M. Sci. Laura Pruzzo
Departamento de Producción Animal, Facultad de Agronomía UBA

Introducción

En la entrega anterior hemos tratado la incertidumbre en forma conceptual. La estadística nos permite abordar cuantitativamente al menos parte de dicha incertidumbre; dándole valores a la fracción de incertidumbre conocida, y describiendo parte de la variabilidad natural. Debemos recordar que la precisión provista por el análisis estadístico no es completa, y siempre existirá en un contexto de riesgo, una parte de incertidumbre que no se puede cuantificar. Pero siempre será mejor evaluar los aspectos conocidos, que ignorarlos.

Como evaluadores, podemos proveer una idea del rango de posibles riesgos, mediante la obtención de estimadores puntuales. Existen situaciones en las que, si la variabilidad e incertidumbre son suficientemente reducidas, un estimador puntual resulta confiable; y, en tanto valor único, también resulta fácil de interpretar. En análisis de riesgo existen estimadores puntuales de particular interés, que se presentan en la siguiente sección. Asimismo, los conceptos básicos previos de Dosis de Referencia, LOAEL y NOAEL, en tanto se informen como valores puntuales promedio o extremos, constituyen también ejemplos de estimadores puntuales utilizados en riesgo ambiental.

1. Estimadores de tendencia central (ETC)

Destinados a brindar una caracterización de las variables de exposición y de riesgo, para un individuo típico de la población (por ejemplo, un consumo promedio de 2 litros de agua por día). Usualmente se basan en la *media aritmética* o estimación promedio, y en la *mediana* de los factores de exposición.

Los ETC interesan particularmente en la evaluación de riesgos ecológicos, caso en el cual interesa el efecto sobre la sustentabilidad de una comunidad, más que en un individuo particular.

2. Estimadores de máxima exposición o Máxima Exposición Razonable

Previstos para estimar el riesgo que se espera que ocurra en el rango superior del conjunto de datos. También considerados como valores *conservadores*. El objetivo es considerar un caso conservador, es decir, *bien por arriba del promedio, pero aún dentro del rango de exposición posible.*

Usualmente se basan en la utilización de *percentiles* (percentil 90, 95 o 99) e *intervalos de confianza*. Cabe destacar que los ETC se acompañan de intervalos de confianza, para tener una apreciación de la incertidumbre. Por ejemplo, en los algoritmos para estimar exposición, se utiliza el límite superior del intervalo de confianza del 95% para la concentración promedio y para la tasa de ingestión. También se utiliza el valor puntual correspondiente al dicho límite para el caso del factor de potencia carcinogénica. Asimismo, los factores de incertidumbre utilizados para obtener la dosis de referencia, pueden verse como estimadores de límites superiores para la incertidumbre en cada paso de extrapolación de los datos toxicológicos disponibles.

3. El enfoque del Análisis de Riesgo por niveles

Actualmente, se considera que la etapa de caracterización del riesgo puede desarrollarse en tres niveles de análisis. El enfoque determinístico o de estimadores puntuales, debe realizarse como primer nivel de análisis. Si las estimaciones superan los niveles críticos, el analista deberá considerar si la información disponible sustenta una decisión de remediación, o si se requiere una evaluación adicional.

Primer Nivel

Se utilizan estimaciones puntuales para representar cada uno de los factores de exposición, con la metodología básica que se describe en esta entrega. Los resultados obtenidos pueden ser suficientes, dependiendo de la magnitud de los estimadores y del nivel de confianza en ellos. La variabilidad en los niveles de exposición, puede evaluarse considerando estimadores de tendencia central y de máxima exposición. La incertidumbre asociada al muestreo se toma en cuenta utilizando los límites del intervalo de confianza del estimador puntual (por ejemplo, la concentración del tóxico en el medio)

Es importante destacar que en algunos casos los resultados del primer nivel pueden alcanzar para la toma de decisiones:
Si los riesgos para un escenario de máxima exposición o muy conservador se hallan por debajo de los valores considerados "de preocupación", entonces no se requerirán mayores análisis.
Recordemos que el uso de valores puntuales puede ignorar parte de variabilidad; no obstante lo cual, son importantes herramientas en el proceso de evaluación del riesgo para identificar situaciones en las cuales, aún los supuestos más conservadores indican un bajo nivel de riesgo.

Segundo Nivel

Es un nivel intermedio de análisis que puede aportar una evaluación más realista de la exposición, o un modelo más detallado. Puede involucrar el *análisis de sensibilidad* para identificar las variables más importantes que afectan la estimación de riesgo y contribuyen a la incertidumbre. Dicho análisis

involucra conducir estudios de caso con distintos valores y observar cambios; si éstos son muy pronunciados, no se tiene la confianza suficiente en la estimación de riesgo, indicando la necesidad de un análisis más refinado.

Tercer Nivel

Se trata de un nivel de análisis *probabilístico,* con el uso de modelos como el proceso Montecarlo (se verá más adelante en el curso), para estimar el impacto relativo de la variabilidad natural y de la falta de información, sobre la incertidumbre general de la estimación de riesgo.

4. Ventajas y desventajas del enfoque determinístico

Como ventajas de este enfoque podemos mencionar la posibilidad de describir la variabilidad mediante una combinación de ETC y valores extremos. Presenta utilidad como método de evaluación de un problema de riesgo ambiental, si los estimadores de máxima exposición resultan mucho mayores o menores al nivel crítico regulatorio. Asimismo presenta facilidad para entender y comunicar los riesgos y una reducida demanda de recursos.

Como desventajas se señala que los estimadores puntuales funcionan como límites fijos y no reflejan o identifican la incertidumbre; tampoco permiten identificar factores clave en la determinación del riesgo. No proveen información acerca de cuál es la probabilidad de que el riesgo exceda un nivel crítico, no se utiliza toda la información disponible y pueden existir inconsistencias entre análisis por el uso de diferentes valores puntuales.

5. Riesgo por exposición a radón: estimadores puntuales de variables de exposición

En los años ochenta, la ocurrencia de exposiciones posiblemente elevadas de los productos de desintegración radiactiva del radón 222, presentes en el aire del interior de las viviendas familiares, condujeron a una evaluación de los datos de concentración de radón en hogares de Estados Unidos (Nero *et al*, 1986). Se ha identificado al radón como un potente carcinógeno, con 1×10^{-8} muertes por cáncer por hora de exposición en 1 pci/l. Un curie es una medida de desintegración por unidad de tiempo; 1 pci = 0,037 desintegraciones por segundo.

La concentración de radón es una de las variables del algoritmo de exposición, pudiendo ser caracterizada mediante estimadores puntuales sobre la base de una muestra representativa de la población. En el estudio citado, se determinó un valor de concentración promedio nacional de 1,5 pci/lt, pero se encontró que el 1% a 2% de los hogares relevados, excedía los 8 pci/lt.

Recordamos que:

Riesgo de cáncer = f(potencia, exposición)

Con el valor promedio de 1,5 pci/lt, el riesgo de cáncer de por vida resultaría en el valor 8,2 x 1E-05. Que utilidad tiene esta estimación? Desde el punto de vista de la toma de decisiones, muy poca, puesto que las viviendas relevadas con las concentraciones máximas, sufren un riesgo mucho mayor. Si casi todos los hogares se encuentran cerca del nivel promedio, las implicancias serán muy diferentes que si el 99% de los hogares presenta valores muy bajos, pero un 1% a 2% tiene niveles más elevados de concentración, y un 0,001%, niveles elevadísimos, por ejemplo 10 pci.

5.1 Medidas de tendencia central y análisis exploratorio de datos

El siguiente cuadro consigna una muestra de los niveles de radón medido en hogares de una localidad.

Cuadro 1. Niveles de radón en las viviendas, en pCi/l

Casa	1	2	3	4	5	6	7	8	9	10	11
Nivel	4,04	4,60	5,73	5,39	2,37	5,39	4,60	5,05	4,38	5,05	4,04

Ejercicio 1. Obtenga la media del nivel de radón de la muestra.

La media muestral es, sin duda, la medida más común de tendencia central. Sin embargo la media y el desvío estándar son sensibles a los valores alejados, y unos pocos puntos extremos pueden distorsionar severamente la estimación. Veremos que la mediana es otro ETC, que es más robusto a valores extremos, también veremos que es un percentil especial. La mediana se calcula ordenando los datos en forma ascendente. Con un número impar de observaciones, la mediana es el valor intermedio. En el caso de un número par de valores, se sigue la convención de definir la mediana como el promedio de los valores de las dos observaciones intermedias.

Eercicio 2. Obtenga la mediana de este conjunto de datos

Aunque la media es la medida de localización central que más se usa, hay casos en que se prefiere la mediana. La media se ve afectada por valores extremadamente pequeños o extremadamente grandes, pudiendo resultar poco representativa; en tales casos, es probable que la mediana cambie muy poco, o nada. Por ejemplo, en estos conjuntos de datos :

- ✓ La mediana de 13,11,17 es 13
- ✓ La mediana de 13, 11, 17, 15 es (13+15)/2 = 14
- ✓ La mediana de 13, 11, 20.234.234, todavía es 13.

En general, podemos decir entonces que si un conjunto de datos tiene valores extremadamente altos o bajos, la mediana es la medida preferida de locación central, debido a su propiedad de resiliencia a los valores extremos o "*outliers*". Esto es especialmente relevante si los outliers representan datos incorrectos. Sin embargo, se debe tener mucho cuidado: si los datos son correctos, podemos tener interés en tales valores. Por ejemplo, la adición de una casa con un nivel de 40 pCi/l a la población en estudio, puede representar una lectura errónea, o una casa con un riesgo excesivamente alto...dejamos como ejercicio adicional 2.b, calcular y comparar la media y la mediana con este dato adicional.

Ejercicio 3. a) Calcule la media y la mediana de la concentración de radón para los datos del cuadro 2. b) Si el nivel crítico de consideración es de 4 pCi/l, porqué la mediana sería problemática en este caso? 2.c) Realice un histograma resumiendo los datos del cuadro 2.

Cuadro 2. Otra muestra de hogares y su nivel de radón, en pCi/l

Casa	1	2	3	4	5	6	7	8	9	10	11
Nivel	3,83	3,54	3,46	3,90	12,06	8,97	9,24	14,67	3,13	3,88	2,99

La situación planteada en 2.b no es muy común, dado que las medidas se aplicarían a los hogares individuales, no por áreas. La media y la mediana capturan solamente un aspecto de los datos, y nada dicen de otro importante aspecto, el agrupamiento en "alto nivel" y "bajo nivel". Dado que los niveles de radón obedecen a varias causas, es posible que todas las casas del grupo alto tengan algo en común, por ejemplo, un tipo particular de lecho rocoso, ausencia de sótanos, etc., lo cual representa información mucho más útil que la media o mediana exclusivamente (ver el histograma).

Una tercera medida de localización es la *moda*, valor de los datos que se presenta con más frecuencia. Si los datos tienen exactamente dos modas, son bimodales; si tienen más de dos modas, son multimodales. En este último caso casi nunca se menciona la moda, porque no ayudaría a describir la localización de los datos.

Percentiles

Un percentil da información acerca de cómo se distribuyen los valores sobre el intervalo, desde el menor hasta el mayor. Para datos que no tienen muchos valores repetidos, el percentil *p* divide los datos en dos partes: el *p* % de las observaciones tienen valores menores que el percentil *p* y aproximadamente el (100- p) % de las observaciones tienen valores mayores que el percentil *p.*

Un ejemplo usual es el de las calificaciones en una prueba de admisión a universidades. Supongamos que un aspirante alcanza una clasificación de 54 puntos. No sabemos cómo fue su desempeño en relación a otros; sin embargo si su calificación corresponde al percentil 70, sabemos que aproximadamente,

70% de los alumnos tuvo calificaciones menores a ese valor y 30% tuvieron calificaciones mayores que él.

Los _cuartiles_ son percentiles específicos. Con frecuencia, se dividen los datos en cuatro partes, cada una con aproximadamente el 25% de las observaciones. A los puntos de división se los denomina cuartiles y se definen:

Q1= primer cuartil o percentil 25
Q2= segundo cuartil o percentil 50, también es la mediana
Q3= tercer cuartil o percentil 75.

Nótese que el intervalo de Q1 a Q3 da la porción media central o 50% intermedio de los datos, y así se define el _rango intercuartil_ como:

RIQ= Q3-Q1

Y es una medida de la dispersión con respecto a la mediana. Sean *a* y *b*, los valores mínimos y máximos de la variable respectivamente. Entonces los cinco parámetros:

1. *Valor mínimo a*
2. *Primer cuartil Q1*
3. *Mediana Q2*
4. *Tercer cuartil Q3*
5. *Valor máximo b*

Son referidos como _Resumen de cinco números_. Juntos, estos parámetros proporcionan una gran cantidad de información acerca de la distribución de los valores de la variable en términos de centro, dispersión y sesgo. Gráficamente, los cinco números a menudo se presentan como un _diagrama de caja_, que consiste en una línea que se extiende desde *a* hasta *b*, con una caja rectangular de Q1 a Q3 y con marcas para ambos límites y para la mediana. Tanto el resumen de cinco números como el diagrama de caja constituyen técnicas de análisis exploratorio de datos.

5.2. Midiendo la dispersión

Si tuviéramos que elegir al azar una casa de los datos del cuadro 1, “esperaríamos” que la concentración fuera la media, es decir, 4,60 pCi/l. Pero esto no será siempre así. La forma más común de pensar si la media representa bien los datos, es pensar cuánto se espera que un dato en particular varíe de la media; esto es lo que describe la varianza. Si un conjunto de datos es una población, la *varianza de la población* es:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Si se tiene solamente una muestra de todos los datos posibles, los valores muestrales se usarán para estimar los valores poblacionales. La *varianza muestral* entonces es:

$$S^2 = \frac{\sum\left(x_i - \overline{X}\right)^2}{n-1}$$

Es importante notar que utilizamos (n-1) en el denominador, porque no conocemos la verdadera media μ. En cambio, la estamos aproximando con X, entonces ocurre una pérdida de independencia, reflejada en la pérdida de un grado de libertad.

Finalmente, más a menudo nos referimos a la *desviación estándar*, que es la raíz cuadrada de la varianza; y que se mide en las mismas unidades que las de los datos originales.

Al usar la media y la desviación estándar podemos determinar la localización relativa de cualquier observación. Llamamos *valor z* a un valor asociado a cada valor x_i de los datos tal que:

$$z_i = \frac{x_i - \overline{X}}{s}$$

Este valor puede interpretarse como el número de desviaciones estándar que x_i dista del promedio. Un valor z igual a cero indicaría que el valor de una observación es igual a la media.

En las aplicaciones prácticas, se ha encontrado que muchos conjuntos de datos tienen una distribución en forma de campana o distribución normal. Si bien en los problemas ambientales, siempre deberemos cuestionarnos la validez de este supuesto, cuando se cumple es muy útil. Por ejemplo, puede aplicarse una *regla empírica* para determinar el porcentaje de elementos que debe estar dentro de una determinada cantidad de desviaciones estándar:

- Aproximadamente 68% de los elementos están a menos de una desviación estándar de la media
- Aproximadamente 95% de los elementos están a menos de dos desviaciones estándar de la media
- Casi todos los elementos están a menos de tres desviaciones estándar de la media

Consecuentemente, si conocemos la media y la desviación estándar podemos tener una buena idea de donde está la mayoría de los datos. La siguiente fórmula nos da el rango dentro del cual esperaríamos encontrar el 95% de los datos:

Rango= { μ – 2σ, μ + 2 σ }

Ejercicio 5. Para los datos del cuadro 1 calcule la varianza, el desvío estándar y el rango.

5.3. *Inferencia: de vuelta a la población*

A pesar del esfuerzo de recolección de los datos, el interés está realmente en la población; las muestras son dispositivos para reunir información acerca de ella. Los valores puntuales obtenidos de las muestras, son estimaciones del valor verdadero, desconocido, de la concentración de radón en la población.

El intervalo de confianza (*IC*) es la herramienta estadística que indica la probabilidad de que un intervalo alrededor de la media muestral comprenda a μ. A menudo se utiliza un IC del 95%, pero en las aplicaciones de riesgo ambiental nos encontraremos muchas veces con la necesidad de tener mucho más que el 95% de confianza en ciertos valores, como por ejemplo la posibilidad de una reacción nuclear incontrolada o un accidente de avión. Un 5%, o aún un 1% de probabilidad de error es demasiado para muchos eventos. En el caso de un IC del 90%, nos interesan tanto el límite inferior como el superior, entonces el nivel de confianza = α/2 = 1 - 0,9/2 = 0,05. Es decir, toleramos un 5% de posibilidad de sobreestimar y un 5% de probabilidad de subestimar la media.
La distribución *t* es utilizada para determinar un IC para la media:

$$IC(1-\alpha) = \overline{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Donde el valor *t* da el área α/2 en el extremo derecho de la distribución t, con n-1 grados de libertad (GL) y coeficiente de confianza (1-α). La distribución t depende de los GL; a medida que éstos aumentan, se acerca a la forma de la distribución normal estándar. Los IC son sensibles a valores extremos, ya que la media y el desvío estándar lo son.

Supongamos ahora que disponemos de un conjunto mayor de datos de concentración de radón; en este caso el archivo contiene información de 43 hogares. Para esta tercera muestra se calcularon los siguientes valores:

Media = 3,85 pCi/l
Mediana = 3,83 pCi/l
Desviación estándar = 1,40 pCi/l

Dado que tenemos 42 GL, el valor t de tablas es 1,68 y podemos calcular el IC 90% como sigue:

IC_{90} = 3,85 +- 1,68 x (1,4/ V43) = [3,49 pCi/l, 4,21 pCi/l]

En otras palabras, hay una confianza del 90% en que el rango [3,49-4,21] pCi/l contiene el verdadero valor. O bien, la verdadera media está entre 3,49 y 4,21 pCi/l con un 90% de confianza.

Deseamos saber ahora, si una muestra representa adecuadamente a la población. El intervalo de confianza describe el rango de posibles valores reales basados en la muestra, en tanto que el <u>*test de hipótesis*</u> nos permite

probar formalmente si un valor, distinto al que hemos calculado, tiene posibilidad de ser o no, el valor verdadero.

Ejercicio 6. La Agencia Ambiental ha establecido que el nivel crítico de concentración es 4 pCi/l, por encima del cual deberán tomarse medidas de remediación. Con los datos de la tercera muestra y un nivel de significación de 0,10:

a) Pruebe la hipótesis de que la media verdadera de la concentración de radón en la población, es igual al nivel crítico (*el test es de dos colas pues nos interesa saber si la media es mayor o menor al valor crítico*).
b) Qué seguridad tiene Ud. de que el verdadero nivel promedio de concentración de radón es inferior a 6 pCi/l?.

El conjunto de datos define solamente la media de los hogares muestreados, por lo tanto hay cierta incertidumbre acerca de la media poblacional. En tanto el IC provee un rango de certidumbre alrededor de la media muestral, expresado como probabilidad o porcentaje, el test de hipótesis nos permite rechazar una hipótesis con una certidumbre probabilística.
El test utilizado a menudo es el de Student. La distribución t de Student tiene “colas” gruesas, es decir, un mayor número de datos o eventos ubicados lejos de la media. Estas colas incorporan la incertidumbre que resulta de tener una muestra y no los datos poblacionales. El estadístico t es:

$$t_{n-1} = \frac{\overline{X - \mu_0}}{s / \sqrt{n}}$$

<u>Errores tipo I y tipo II</u>

Si rechazamos una hipótesis que en realidad es verdadera, ocurre un error de tipo I. Si aceptamos una hipótesis falsa, ocurre un error de tipo II. Los errores tipo I pueden ocurrir por seleccionar un modelo o hipótesis inapropiada, o por poner demasiada confianza en datos insuficientes o no representativos.
Muchos científicos que estudian la atmósfera, están de acuerdo en que la temperatura global se ha incrementado 0,4 a 0,6 grados en los últimos cien años. Atribuir esto o aceptar que se debe a la variación climática natural, es un error de tipo II (aceptar una hipótesis falsa), si la verdad es que este efecto ha sido inducido por acción del hombre.
Desafortunadamente hay cierto compromiso entre ambos errores. Si elegimos tener mayor certeza de no rechazar hipótesis verdaderas, aumenta la posibilidad de aceptar hipótesis falsas. Debemos tener en cuenta que, aún en una herramienta analítica existen juicios de valor acerca de donde establecer los niveles de confianza. Recordemos que acertar un 95% de las veces, significa equivocarse una de cada veinte veces.

Bibliografía

Anderson, D., Sweeney,D. y Williams, T. 2004. Estadística para administración y economía. International Thomson Editores.

Kammen,D. y Hassenzahl,D. 2001. Should we risk it? Ch. 3: Statistics for Risk Analysis. Princeton University Press.

Nero, A., M. Schwer, W. Nazaroff y K. Revzan. 1986. Distribution of airborne radon 222 concentration in US homes. Science, Vol 234 issue 4779. Pags. 992 a 997

Schwarz, C. 2007. Sampling, experimental design and analysis for environmental scientists. Simon Fraser University

US EPA. 2004. Air toxic risk assessment manual.

Trabajo Práctico 3

Uso de estimadores puntuales en Análisis de Riesgo Ambiental

A) Resolver los ejercicios de la entrega

B) Desarrollar el Estudio de caso: DDT en aves silvestres

1) Con el archivo de datos suministrado en planilla Excel, realice el análisis determinístico, utilizando estimadores puntuales e histograma. ¿Qué concluye, teniendo en cuenta el nivel crítico para la especie de 98 ppm?

2) Con el resumen de cinco números, confeccione el diagrama de caja. ¿Hay valores alejados?

3) Con los resultados de 1) y 2), considera Ud. suficiente este nivel de análisis para el manejo del riesgo?

4) Otro grupo de científicos tomó la una nueva muestra de 10 aves ; las mediciones de DDT fueron las siguientes:

   100 -105 – 97 – 103 – 96 – 106 – 102 – 97 – 99 – 103

Para esta muestra obtenga el IC 95% para la media. ¿Qué concluye, teniendo en cuenta el nivel crítico para la especie de 98 ppm?

5) Con la nueva muestra en 4) y un nivel de significación del 5%, hay evidencia de que el nivel medio de DDT para toda la población es distinto de 98 ppm?